Could Reliability Naturally imply Safety?

Could Reliability Naturally Imply Safety?

Spyridon Orestis Palermos

ABSTRACT. The aim of the present paper is to argue that robust virtue epistemology is correct. That is, a complete account of knowledge is not in need for an additional modal criterion in order to account for knowledge-undermining epistemic luck. I begin by presenting the problems facing robust virtue epistemology by examining two prominent counterexamples—the Barney and 'epistemic twin earth' cases. After proposing a way in which virtue epistemology can explain away these two problematic cases, thereby, implying that cognitive abilities are also safe, I offer a naturalistic explanation in support of this last claim, inspired by evolutionary epistemology. Finally, I argue that naturalized epistemology should not be thought of as being exclusively descriptive. On the contrary, the evolutionary story I offer in support of the claim that reliability implies safety can provide us with a plausible epistemic norm.

Abilities in Epistemology

One of the most promising intuitions in contemporary epistemology is the idea that knowledge is accurate (i.e., true) belief, which is the product of *cognitive ability*. Call this the *ability intuition* on knowledge. Understood as both a necessary and sufficient condition on knowledge, the aforementioned intuition is captured by robust virtue epistemology in ways close to the following one:

> *S* knows that *p* if and only if *S* believes the truth regarding *p* because of (or out of, or through) a reliable belief-forming process (i.e., a cognitive ability).

Similar formulations to the above can be found in the writings of such key figures of contemporary virtue epistemology as John Greco and Ernest Sosa.[i] This is a *robust* version of virtue epistemology as it is supposed to capture all the qualifications for knowledge, including the intuition that knowledge excludes luck, without the need for an additional anti-luck condition, such as the safety principle.[ii]

To see why robust virtue epistemologists might be so confident, consider the well-known (due to Chisholm) Roddy case:

> Roddy (Pritchard 2009: 11)[iii]
>
> Roddy is a farmer. One day he is looking into a field near-by and clearly sees something that looks just like a sheep. Consequently he forms a belief that there is a sheep in the field. Moreover, this belief is true in that there is a sheep in the field in question. However, what Roddy is looking at is not a sheep, but rather a big hairy dog that looks just like a sheep and which is obscuring from view the sheep standing just behind.

As all the typical Gettier cases, the above example demonstrates that even though Roddy is justified in holding his belief (he sees a sheep-shaped object) *and* his belief is true (there is a sheep behind the sheep-shaped object), we must disallow knowledge to Roddy because of the luck involved in the way he gets things right. Robust virtue epistemologists, however, can accommodate this counterexample to the 'justified-true-belief' account of knowledge, by claiming that, contra to their account, Roddy believes the truth of the matter because of/through/out of luck; the content of his belief may still be the product of a cognitive ability, but his believing the truth is due to luck. Put another way, Roddy believes from ability and he believes the truth, but the fact that he believes from ability does not explain why Roddy has a true belief; luck does.

Robust virtue epistemology, therefore, is considered to be an adequate account of knowledge, one that can handle the knowledge-undermining epistemic luck involved in typical Gettier-cases. Present in the literature, however, there are two more counterexamples that allegedly place robust virtue epistemology in a rather awkward position. Let us consider them in a chronological order of appearance.

First comes the Barney-case:

Barney (Pritchard 2009: 12)[iv]

> Barney is driving through the country and happens to look out of the window into a field. In doing so, he gets to have a good look at a barn-shaped object, whereupon he forms the belief that there is a barn in the field. This belief is true, since what he is looking at is really a barn. Unbeknownst to Barney, however, he is presently in 'barn-façade country' where every object that looks like a barn is a convincing fake. Had Barney looked at one of the fake barns, then he would not have noticed the difference. Quite by chance, however, Barney just happened to look at the one real barn in the vicinity.

Barney comes to truly believe that he is looking at a real barn by employing his cognitive abilities; he is looking directly at the barn. Therefore, this is supposed to be a case where despite the fact that Barney's true believing is because of/out of/through his cognitive ability, his cognitive success cannot be called knowledge given that his belief could have so easily been false (Barney is in a barn-façade environment).

Before moving on to the next counterexample, notice a subtle difference. In regular Gettier-cases like the Roddy counterexample, the knowledge-undermining luck is very direct in that the luck concerns the relation between the belief and the fact. One's belief is erroneously formed but there is a lucky fact that renders it true.

Luck intervenes between the belief and the fact. On the contrary, in cases like the one Barney is in, the knowledge-undermining luck is quite indirect; indeed, it is specifically *environmental*. Barney does look at a real barn and believes that there is a real barn in front of him. At a first glance, there is nothing wrong with the way he forms his belief. Given, however, the particular *environmental conditions,* he cannot acquire knowledge in this way. Luck interferes with the environment in such a way that even a normally well-formed belief will be a lucky one, if true. So, it seems that while robust virtue epistemology can deal with the normal (i.e., non-environmental) knowledge-undermining *epistemic* luck involved in Gettier cases, it may not be able to deal specifically with *environmental* knowledge-undermining luck, involved in Gettier-*style* cases such as the one Barney finds himself in.

And neither, it seems, can it deal with 'epistemic twin earth' counterexamples. Drawing on Putnam's (1975) famous thought experiment, which (supposedly) shows that the contents of our beliefs cannot be narrowly defined, Kallestrup & Pritchard (*forthcoming*) have put forward an epistemic analogue to the effect that knowledge ascriptions cannot solely depend on psychological factors—i.e., whether an agent possesses a cognitive ability—because they also heavily rely on the *safety* of the environment the epistemic agent is situated in. In other words, robust virtue epistemology cannot fully account for knowledge because an adequate account requires the explicit endorsement of a condition that will ensure one's beliefs are safely formed.

Before turning to the actual example, however, we should first go through some of the technical terms that Kallestrup & Pritchard use in their exposition of the thought experiment (*forthcoming*: 6):

- Call the place where the subject is currently situated in the *local environment*:

  It contains the objects and properties that are the proximate causes of her current perceptual experiences. Take facts to be objects instantiating a property at a time. If the subject now perceives that *p*, then the fact that *p* (the '*p*-fact') is one that concerns her local environment—i.e., it is a *local fact*.

- The *regional environment* is not the place where the subject is currently in, nor the place she is typically located:

  Still, it contains the objects and properties with which she might easily have been causally connected. If the *q*-fact is such that if the subject had not now perceived that *p* then she would have perceived that *q*, then the *q*-fact is one that pertains to her regional environment—i.e., it is a *regional fact*.

> Regional facts, therefore, are close nearby perceptual possibilities, even though, as a matter of fact, they play no causal role in producing the subject's current perceptual experience on which she bases her belief that *p*.

- Finally, the place where the subject is typically, though not at present, located is called the *global environment*:

> It contains the objects and properties with which she ordinarily causally interacts. The *global facts* thus comprise all the facts that extend in space-time beyond the regional facts. Assuming the subject now perceives the local fact that *p*, the fact that *r* is a global fact only if she would not have perceived that *r* had she not perceived that *p*.

Thus understood, one's local environment is the part of one's regional environment that one is currently attending to. One's regional environment (including one's local environment), however, might or might not be part of one's global environment.

We can now proceed with the actual thought experiment. *S* is on earth where all watery stuff is H_2O. Moreover, *S*'s perceptual ability is highly reliable in that a high frequency of her perceptual beliefs are actually true; based on a perceptual experience as of water, *S* forms the demonstrative belief that that's water and intuition dictates that she also *knows* this.

Now, on twin earth, there is an intrinsic physical duplicate of *S*, twin-*S*. On twin-*S*'s global environment, all watery stuff is again H_2O. Accordingly, twin-*S* can reliably form true beliefs about water while situated in her global environment. Moreover, in her local environment, twin-*S*'s belief about water is true as she is currently interacting with H_2O. However, unbeknownst to twin-*S*, her present regional environment is full of twin-water, which is perceptually indistinguishable from water, but with chemical structure XYZ. Therefore, her current water-belief is only luckily true given that she could have so very easily interacted with twin-water, and her belief would have thereby been false: 'that is to say, very easily could twin-*S* have believed that that's water on the same basis—a perceptual experience as of water—without that being so' (Kallestrup &Pritchard *forthcoming*: 7). Given that knowledge excludes luck, it follows that twin-*S lacks* knowledge.

Remember that *S* and twin-*S* are physical intrinsic duplicates embedded in identical global environments, and therefore, on Kallestrup & Pritchard's account, both have the ability to detect water. Accordingly, the upshot of the thought experiment is that their epistemic difference cannot be explained away in terms of the

possession or absence of the relevant cognitive ability. There must be another condition on knowledge that accounts for the fact that *S* knows that *p*, whereas twin-*S* does not.

Accordingly, on the basis of the Barney and 'epistemic twin earth' thought experiments, the authors conclude that robust virtue epistemology cannot ultimately do all the work that is expected to do. Despite initial expectations, the ability intuition on knowledge seems unable to *fully* accommodate the equally important intuition that knowledge must not be due to luck, *viz.* the anti-luck intuition on knowledge. The moral appears to be that these two intuitions about knowledge 'impose independent epistemic demands on our theory of knowledge' (Pritchard *forthcoming*: 2).

Consequently, friends of anti-luck epistemology claim that any adequate theory of knowledge must explicitly have as a central component an anti-luck epistemic condition such as the safety or the sensitivity principle.[v] These modal conditions on knowledge—in contrast to the ability condition on knowledge, which addresses the problem posed by knowledge-undermining luck only indirectly—are primarily targeted to capture the anti-luck requirement, as they are explicitly concerned with the responsiveness of one's belief to relevant counterfactual circumstances (such as the scenario in which Barney looks at a barn-façade instead of a real barn). So the upshot is that for virtue reliabilism to be a fully adequate account of knowledge, it may well have to be supplemented by a specific anti-luck condition on knowledge such as safety or sensitivity.[vi]

A Response

The problem robust virtue epistemology faces has now been sufficiently described. It is time, then, to explore in what ways virtue epistemologists might respond. To start with, we must first provide an account of abilities we can work on.

In general, ability can be thought of as a (cognitive) process, which is reliable. In some more detail, the ability to *R* can be thought of as *reliable* dispositional, or habitual (cognitive) process with outcome *R*, possessed by an agent. Reliability is here understood as a successful track record to *R*. Naturally, the success of a process depends on the 'background conditions' against which the process is exercised.

Notice that 'background conditions' might refer both to a broad *environment*, *E*—a set of relatively stable circumstances—and more particular *conditions*, *C*—a

range of shifting circumstances within *E*. Even though, as Greco (2010: 77) notes, the two terms might overlap this should not create any problems for the account. To provide a few examples, the *environment* might refer to the presence of gravity, air, or good lighting, whereas *conditions* might refer to the absence of oil or ice on the tarmac, or to the requirement that the piano is well tuned. Moreover, *conditions* might refer to the quality of the potential input (i.e., stimuli) that might be fed in the target process, or even to internal aspects of the subject such as sobriety or tiredness.

Having the above considerations in mind, it can now be claimed that a (cognitive) process *P* will signify ability to *R* if and only if its success rate within *E*, *C* to *R* is such that the process counts as reliable. Schematically:

> A (cognitive) process *P* signifies ability to *R* in E, C if and only if S_p (E, C) to *R* > threshold of success rate for reliability.

Notice, further, that abilities, seen as reliable (cognitive) processes, are acquired over a long period of time through constant interaction of the agent with the environment wherein he is typically situated.[vii] In Kallestrup and Pritchard's terms, abilities are tied to the *global environment* in which they are acquired and possessed: 'Abilities are relative only to the stable environment in which whoever has them is typically located […] These are the normal circumstances in which abilities are acquired through learning and sustained through practice' (*forthcoming*: 10). In particular, it could be claimed that cognitive processes are developed both phylogenetically and ontogenetically within a given environment. This is the environment, which the relevant processes evolved/were designed to reliably operate in. In addition, it should be uncontroversial to claim that when some ability is acquired, a physical substrate that supports the relevant process is also developed. This physical substrate will usually be found within the neural architecture of the agent's brain, and possibly her bodily structure as well (I will return to these claims in the last section).[viii] Overall, then, we might say that when in her normal environment a subject exhibits a high success rate to *R* on the basis of a (cognitive) process, the subject *possesses* the ability to *R* (relative to her normal environment).

Given changes in *E*, *C*, however, what the above formulation also demonstrates is that the success rate of an *otherwise reliable* process might significantly fall. Accordingly, given defective *E'*, *C'* one of the following might be claimed:

1) The subject *does not posses* the ability in question any more.

Or,

2) The subject *cannot manifest* the ability in question any more.

Notice this is exactly the situation Barney is in; he is in an abnormal environment—full of barn-facades—wherein his otherwise reliable process of detecting real barns through vision could have so very easily outputted a false belief. Since virtue epistemologists want to avoid using the safety principle—which would obviously offer the desired result—the only way to deny knowledge to Barney, as intuition dictates, is by taking up either option 1) or 2).

Let us explore the first avenue first. This is actually the tactics Allan Millar (2010) employs. Millar first distinguishes between a narrow and a broad construal of abilities. Briefly, to say that abilities are narrowly construed is to follow the typical way of thinking about abilities as having their bases mostly within the organismic boundaries of the individuals, and letting 'environment' and 'conditions' only determine whether they can be appropriately employed. On the contrary, to say that abilities are broad is to support the radical claim that the bases of abilities also include aspects of the 'environment' and/or the 'conditions'. In effect, under the broad construal, abilities are only *possessed* in environments wherein they can be successfully exercised.[ix]

For instance, Millar—embracing the broad conception—holds that perception does not merely supervene on one's psychology but also on physical features of things in the world. Whether a subject *possesses* the ability to detect an object heavily depends on the correlation between the object having the relevant appearance and the object actually being the thing it appears to be. Since this is not the case in the fake barn territory, Millar deals with Barney in the following way:

> When Barney judges falsely in fake-barn territory he fails to exercise an ability to tell of certain structures that they are barns from the way they look. Indeed, he does not have the ability to tell structures around there that they are barns from the way they look. Of course, when he is there he does something like that he also does back home —judge of some structures that look like barns that they are barns—and in doing so he will sometimes judge correctly. But that does not amount to his being able to tell of the structures that they are barns from the way they look (op. cit: 126-27).

According to Millar, then, Barney does not have the ability in question and this allows robust virtue epistemologists to do away with the Barney case: just as

intuition dictates, Barney lacks knowledge simply because, given Barney's abnormal environment, he *does not possess* the ability to detect real barns.

The 'epistemic twin earth' thought experiment, however, comes to the rescue for opponents of robust virtue epistemology. For since *S* and twin-*S* share identical global environments, it is not open for Millar to claim that the ability to detect water is possessed by *S* but not twin-*S*; in both *S* and twin-*S*'s global environments the circumstances are conducive to the development and appropriate employment of the relevant ability and, thereby, they are both capable of detecting water. Therefore, it appears that Millar cannot account for twin-*S*'s epistemic difference in the same way he does for Barney.

At his point, however, Millar could apparently object that Kallestrup & Pritchard beg the question against his view. Why should it be the case that *global,* instead of *regional environments*—as in effect Millar argues—bear on considerations regarding the possession of ability? On a first view, the reason why Kallestrup & Pritchard's account appears to be more appealing is that Millar's account leads to the counterintuitive result that the possession of abilities comes and goes with changes in the epistemic agent's environment; even if I cannot ride my motorbike on sand, I do not thereby need to re-acquire my driving ability when I get back on the street. This would be a rather awkward conclusion to draw. Accordingly, I want to suggest that the reason why Millar's view appears to be counterintuitive has to do with an observation we made earlier. When one acquires some ability on the basis of one's interaction with one's normal (i.e., *global*) environment, this is mirrored in one's neural and bodily architecture. That is, through training within one's normal environment, one acquires a bodily process, which is successful within that particular environment. When one, however, finds oneself in defective circumstances one does not lose the relevant process that the corresponding ability is identified with; the neural and bodily structure cannot—at least not out of a sudden—disappear. What happens, instead, is that the relevant process cannot manifest the reliability it usually does relative to the environment it was acquired for. In other words, given that the physical processes supporting the relevant ability—i.e., the realization basis of the ability—cannot suddenly vanish, what we should claim is that, in *in*appropriate environments, the relevant cognitive processes simply cannot manifest their reliability; therefore, *ability cannot be manifested*.[x] This should also explain why

provided that one moves back to one's normal environment soon enough, one does not need to re-acquire the ability in question.

Accordingly, it appears more appropriate to say that once one has acquired some ability (i.e., has developed the relevant reliable cognitive process) within and relative to one's (*global*) environment, one possesses the target ability wherever one might be situated in, but can only *manifest* it when the appropriate circumstances obtain. Kallestrup & Pritchard concede this point when they write: 'Temporary abnormal environments cannot rob a subject of an ability that she otherwise reliably manifests in the normal run of things' (*forthcoming*: 11).

Notice that this is actually the second option left open for robust virtue epistemologists in order to account for both the Barney and the 'epistemic twin earth' case. Instead of claiming that the agents do not possess the relevant abilities, one should claim that they *do* possess the relevant abilities when considered in relation to their global environments wherein they acquired them and typically employ them, but the special circumstances obtaining in their *regional environments* disallow both Barney and twin-*S*'s abilities to be appropriately *manifested*. In other words, the fact that barn- and water-appearances are not distinctive of real barns and water respectively, is one of the background conditions that need to, but do not obtain for their abilities to be appropriately employed (i.e., manifested).[xi] Accordingly, since no abilities are manifested in those cases, what e*xplains* the two agents' true beliefs is the good luck that they happened to interact with the only real things they were looking for.

As a matter of fact, Kallestrup and Pritchard anticipate this kind of response when they write the following:

> The causal explanatory reading [of robust virtue epistemology] doesn't fare any better on this score. For given that the local environments that *S* and twin-*S* inhabit are completely identical, it is hard to see how there should be a different causal explanation offered of their respective cognitive successes. Since the *explanandum* (i.e., true belief) is identical in the two cases, the explanations would differ only if a difference in *explanans* could be established. But what could that possibly be? Remember that, at this juncture, we take *S* and twin-*S* to be sharing all pertinent cognitive abilities. For while Greco [2009, 21-22] is surely right that abilities are tied to conditions understood as local environments, it remains that only the outcomes of exercising abilities can shift with variations in local facts, and not whether abilities are possessed at all. […] So, if one such ability features in the causal explanation of *S*'s true belief, it would also do so in the case of twin-*S*'s true belief. Likewise, no local physical facts could be cited in one causal explanation but not the other. True, causal explanations frequently involve extrinsic properties, but as *S* and twin-*S* also share global physical environments, Greco needs to show that the differences in their

> regional environments somehow is bound to bear on the causal explanations in question.

Kallestrup and Pritchard are surely correct when they write that twin-*S*'s global environment explains why twin-*S* possesses the relevant ability and that his local environment explains why his belief is actually true. Having conceded this, however, let me also note that the rest of their assessment is not entirely correct. For while the fact that *S* and twin-*S* share identical global and local environments explains why they both believe from ability and that their beliefs are also true respectively, it is only in *S*'s case that robust virtue epistemologists' crucially stronger requirement that 'belief be *true* in virtue of resulting from the exercise of ability' is satisfied. On the contrary, both in the Barney and the twin-*S* case, the overarching factor in the causal explanation of how they believe the truth of the matter is luck, because, given their regional environments, their abilities cannot be appropriately manifested. Of course, this is not to say that their abilities played no role in forming their beliefs, since it is those very abilities that provided them with the content of their beliefs. Crucially, however, given the regional circumstances, their abilities are not the reason why they got to the *truth* of the matter.

So, what in effect I am arguing for is that the manifestation of abilities depends on one's *regional environment*. Accordingly, the differences between Barney's and twin-*S*'s *regional environments* and the regional environments of their counterparts should adequately explain their epistemic differences. That is, as Kallestrup & Pritchard request at the end of the above quote, we now have a reason to think why the differences in the epistemic agents' *regional environments* are bound to bear on the causal explanations regarding their epistemic differences. One's global environment explains whether one possesses some ability and one's local environment explains whether one's ability is actually accurate. However, one's believing the *truth* should be because of the manifestation of the relevant ability, which, in turn, depends on one's *regional environment.*

Before concluding, however, here are two possible ways in which Kallestrup & Pritchard could possibly rejoin the discussion. First, they could claim that it is facts about one's local environment that should bear on considerations regarding the manifestation of some ability. To see, however, why this is a non-starter first recall that whether an already possessed ability is manifested within a specific environment

depends on the (actual or theoretical) success rate of the corresponding cognitive process within that environment. Notice, however, that one's *local environment*, as defined by Kallestrup & Pritchard, picks out only a particular, well-defined and non-changing aspect of the world against which no success rate can be calculated simply because there is only one possibility available. On the contrary, one's *regional environment* offers the set of relevantly close possibilities for interaction, on the basis of which success rates can be appropriately estimated, thereby allowing the success rate of an otherwise reliable process to be assessed.[xii]

Second, another possible objection could be to claim that abilities are dispositions. Regional facts, however, do not disallow dispositions from being manifested. For example, salt is soluble in water, but not XYZ. Accordingly, if we drop crystals of salt in a glass of water, it will dissolve even if many other glasses, which are full of XYZ, surround that particular glass. Given this discrepancy, therefore, there must be a problem with the offered account of abilities. In response, notice the equivocality involved in the meaning of the word 'disposition'. On one hand, to claim that cognitive abilities are dispositions means that abilities are character traits, or habitual behaviors that the agent tends to exhibit. On the other hand, to claim that an object, or an entity has a certain disposition to behave in a certain way, such as that the vase is fragile or that salt is soluble in water is quite a different claim, whose meaning is also the topic of a very thorough and presently unsettled debate.[xiii] An indication—not irrelevant to the present discussion—that the two uses of the word are not the same is the fact that dispositions, in the first sense of the term, can only be acquired and sustained through practice, whereas dispositions, in the second meaning of the term, can be possessed by an entity even if they are never actually manifested (e.g., a vase may be fragile even if it has never been broken). Accordingly, given both that there is an ambiguity in the term, and that the second sense of the term is far from well-understood, the fact that cognitive abilities are dispositions that seem to behave differently from the dispositions possessed by objects is not compelling grounds for rejecting the offered account of abilities.

To conclude, then, we now have a principled way to account for the epistemic difference between *S* and twin-*S,* and why Barney lacks knowledge without appealing to an anti-luck condition on knowledge. That is, abilities understood as reliable cognitive processes whose manifestation depends on one's regional environment provide robust virtue epistemology with a principled way to account not just for

epistemic, but *environmental* knowledge-undermining luck as well. Should this, however, come as a surprise, or is there a natural explanation for the fact that abilities appear to be *safe*?

Reliability and Safety, Naturalized

It is not long ago that, in a visiting lecture, Dan Dennett mentioned 'Darwin's strange inversion of reasoning' (Dennett 2009): the idea that we can have competence without comprehension, or, in other words, that it is not necessary to think that every purposeful object must have a designer. Evolutionary biology demonstrates that all living species exhibit traits and/or abilities that help them deal with the natural environments they inhabit. Crucially, however, far from being the products of a meaningful design, the evolution of the relevant traits is the product of a meaningless process of trial-and-error that extends thousands of years in the past of each species' lineage.

There is no doubt that human beings too, as natural beings, are the products of evolutionary development. Accordingly, their capacities for knowledge and belief are also the products of a natural evolutionary development. This, in turn, provides a basis for thinking that knowing, as a natural activity, could appropriately be analyzed along the lines compatible with its status, i.e., by the methods of natural science, and specifically evolutionary biology.

This is the project of evolutionary epistemology, which properly falls under the trend of naturalized epistemology. Interestingly, virtue epistemology with its appeal to cognitive abilities is well suited for a naturalized analysis of knowledge; as noted in the previous section, abilities have their realization bases on processes mirrored by the neural and bodily architecture of the organisms that exhibit them. And as Patricia Churchland (1987: 548-49) notes, 'the principal chore of nervous system is to get the body parts where they should be in order that the organism may survive'.

Since it is well beyond the scope of the present paper to offer a detailed account of the evolution of epistemological mechanisms, I will only provide a description in broad strokes. Still, this should hopefully be sufficient to hint at the claim that reliable cognitive processes are also *safe* in the epistemologically relevant sense.

The central point we can start with is an observation made in the previous section, namely that biological development involves both ontogenetic and phylogenetic considerations. That is, the possession of specific traits can be viewed both from the point of view of the development of that trait in individual organisms (ontogeny) and the evolution of that trait in the species' lineage (phylogeny). Although the two levels of adaptation cannot be properly disentangled—since there is a constant mutual interaction between them—it could be claimed that phylogeny is responsible for the carving of the organisms' bodily characteristics that can support a range of abilities, whereas ontogeny is responsible for the more fine grained articulation of the bodies that organisms are born with, throughout the individuals' lifetime.

Crucially, what both levels of adaptation have in common is that they operate on the basis of some kind of trial-and-error procedure of fine-tuning. Phylogenetically, only the organisms gifted with the DNA sequences that will produce the corresponding fittest phenotypes (always relative to a particular environment) will survive to the age of reproduction, and so only the fittest DNA sequences will be passed on to the next generations. This is, however, not an act of meaningful design. Far from it, it is an iterated process of trial-and-error whereby random mutations are selected by the forces of nature.[xiv]

Similarly, most connectionist models of the ontogenetic shaping of neural systems—largely thought to be the most biologically realistic ones—employ *learning* algorithms, with 'backward propagation of errors' being the most typical one:

> 'A connectionist model is characterized by three architectural elements: (1) processing units, (2) connections between processing units, and (3) weights, which are differential strengths of connections between processing units […] The network does not have any initial or built-in organization for processing the input […] All such organization emerges during the training period. The values of the weights are determined by using the learning algorithm called "back-propagation of error." The strategy exploits the calculated error between the *actual* values of the processing units in the output layer and desired values, which are provided by a training signal. The error signal is propagated from the layer backward to the input layer and is used to adjust each weight in the network. The network learns as the weights are changed to minimize the mean squared error […] Thus, the system can be characterized as following a path in weight space until it finds an error minimum' (Churchland 1987: 550).[xv]

So, we see that the strategy for shaping connectionist networks on the basis of 'back-propagation of error' is, again, an iterated process of trial-and-error fine-tuning,

whereby the network tries out several connections for processing the input, so as to eventually minimize error.

This iterated process of fine-tuning by trial-and-error, present in both levels of adaptation, is crucial to see why the reliability of evolutionary epistemic mechanisms might imply safety. Trial-and-error fine-tuning requires a vast number of natural forces (in the case of natural selection) and inputs (in the case of ontogenetic development) in order to bring about the desired results. In nature, these natural forces and inputs originate from one's *global environment* and, crucially, it is the sum of these inputs that determine the sort of epistemic mechanisms that will eventually evolve and develop. What this means, then, is that evolutionary mechanisms are highly calibrated to operate in these environments such that they can accommodate and overcome any physical discrepancies that might occur. As a result, given that they are employed within the environments they were selected and developed for, there should be no defective conditions these epistemic mechanisms will not have been calibrated to compensate for. Consequently, if they operate within their normal environment, evolutionary mechanisms cannot easily produce false outputs. Conversely, if a species' *global* environment suddenly contained such discrepancies as in Barney or Twin-*S*'s *regional* environments, then either the individuals bearing the defective traits (relative to their new environments) would disappear, taking with them the relevant traits, or, if biologically possible, mechanisms that could accommodate the relevant discrepancies would evolve and/or develop.

It follows, then, that so long as a subject employs her abilities in her *global environment*, there are no close possibilities of error; it is not the case that her abilities could have easily been wrong. That is, the products of abilities—defined as reliable cognitive processes—are safe, provided that the relevant abilities are employed within one's *global environment*.

It, therefore, appears that we now have a natural explanation in support of the claim that reliability implies safety. Before closing, however, a few methodological remarks are in order. Obviously, the 'competence without comprehension' idea regarding the evolution and development of cognitive abilities is a strong indication towards the validity of externalist theories of knowledge; one can be competent in performing *R* without having access to (i.e., comprehending) the reasons on the basis of which one can perform *R*.

While this aspect of the analysis might be considered as a welcome outcome

by many contemporary epistemologists, a common objection is that naturalized epistemology cannot account for the normative aspect of epistemology. As Patricia Churchland (1987: 546) comments, however, 'once we understand what reasoning is, we can begin to figure out what reasoning *well* is'. So, do the above considerations point to what *reasoning well* is?

It appears they do. Robust virtue epistemologists' claim is that knowledge is belief, which is true in virtue of cognitive ability (i.e., a reliable cognitive belief-forming process). In addition, the above analysis demonstrates that already possessed abilities can be manifested (i.e., appropriately employed) only when they operate within the environments they were developed and selected for, or relevantly similar ones. Accordingly, robust virtue epistemologists could claim that *S* knows that *p* if and only if *S* believes the truth regarding *p* because of cognitive ability *and S* is situated within her *global environment.*[xvi] In other words, a good epistemic agent *ought to* employ her cognitive abilities and accept the corresponding beliefs only when situated within her *global environment.* Notice, moreover, that one can be very easily aware whether one is situated within one's normal environment. So, for instance, if I were the first to land on Pandora and wanted to believe what is true—which I probably would for reasons of survival—I would be a very bad epistemic agent were I to trust my otherwise reliable senses. In that planet, what looks and feels like water might not be water and plants, for all I know, might have feet and teeth. This is not my normal environment.

Now, the above considerations bring to mind Greco's subjective justification requirement for knowledge. To see why, let us follow Greco who notes (1999: 285) that 'it is not enough that one's belief is formed in a way that is objectively reliable; one's belief must be formed in a way that is subjectively appropriate as well'. Nevertheless, in order to remain fast to externalism Greco warns that subjective justification must be accommodated in a way that does not involve knowledge of, or even beliefs about reliability. Accordingly, he proposes (1999: 289) that 'a belief *p* is subjectively justified for a person *S* (in the sense relevant for having knowledge) if and only if *S*'s believing *p* is grounded in the cognitive dispositions that *S* manifests when *S* is thinking conscientiously' (i.e., when *S* is motivated to believe what is true). Now, according to the normative claim offered in the previous paragraph, a conscientious thinker, i.e., one who wants to believe what is true, is one who is sensitive to the aptness of one's environment. Crucially to the point, however, notice

that epistemic agents do not need to have any positive beliefs that their belief-forming processes are appropriate relative to the environment they employ them. It just happens, as in the case of landing on Pandora, that if the environment appears to be hostile or, at least, a new one, then (and only then) conscientious thinkers will just proceed with caution, being suspicious of the deliverances of their cognitive abilities.[xvii]

Conclusion

In summary, we have seen a way in which robust virtue epistemology might explain away troubling counterexamples such as the Barney and the 'twin epistemic earth' cases. In such cases, while the normality of the agents' *global* and *local* environments explains why they believe from ability and why their beliefs turn out to be actually true respectively, the abnormality of their *regional environments* disallows them to manifest their cognitive abilities. Accordingly, their beliefs are true not because of their cognitive abilities but in virtue of luck. As far as those two counterexamples are concerned then, virtue epistemology does not appear to be in need of being supplemented by a distinct anti-luck condition on knowledge. Evolutionary and developmental considerations, moreover, appear to suggest that reliable belief-forming processes are also safe in the epistemologically relevant sense. The reason is that cognitive abilities, being the products of natural selection and ontogenetic development, are calibrated to accommodate possible discrepancies within the environment that shaped them. Accordingly, given they are exercised within one's global environment, cognitive abilities cannot easily be wrong. That is, there are no close nearby possibilities in which they could produce false outputs; they are safe. Interestingly, these considerations do not just offer a descriptive analysis of knowledge but they also provide an instructive epistemic norm. That is, epistemic agents are objectively justified when their beliefs are indeed the products of reliable belief-forming processes; epistemic agents, however, in order to be also subjectively justified, ought to accept the deliverances of their cognitive abilities only when they *don't have* any doubts that they operate within their normal environments and under the appropriate conditions (or relevantly similar ones).[xviiixix]

Spyridon Orestis Palermos
University of Edinburgh
Philosophy, School of Philosophy, Psychology and Language Sciences
Dugald Stewart Building
3 Charles Street, George Square
Edinburgh EH8 9AD
S.O.Palermos@ed.ac.uk

Notes

[i] See (Greco 2010) and (Sosa 2007). Note, however, that both of these authors add a subjective condition on knowledge, as well. Sosa holds that reflective knowledge is apt belief, aptly formed. Greco, in turn, holds that whereas a belief is objectively justified for *S* if it is the product of a reliable belief-forming process, *S* won't have acquired knowledge unless he is also subjectively justified by being conscientious, i.e., by accepting the beliefs, which are the product of the belief-forming processes that he manifests when he is motivated to believe what is true. I will return to Greco's subjective justification criterion at the end of the last section.

[ii] See ft. 5 for a formulation of the safety principle.

[iii] The Roddy case is described in Chisholm (1977: 105).

[iv] The Barney case is described in Goldman (1976) and credited to Carl Ginet.

[v] The *sensitivity principle* is usually formulated as follows: If *S* knows that *p,* then *S's* true belief that *p,* is such that, had *p* been false, *S* would not have believed *p.* The classic defenses of the sensitivity principle can be found in Dretske (1970) and Nozick (1981). The *safety principle* is usually understood thusly: if *S* knows that *p*, then *S's* true belief that *p,* is such that *S's* belief that *p* could not have easily been false. For recent defenses of the safety principle see Sosa (1999, 2000) and Pritchard (2002, 2008). For a very good discussion concerning the relation between the ability and the anti-luck intuition on knowledge see Pritchard (*forthcoming*).

[vi] Consider for example *Anti-Luck Virtue Epistemology: S knows that p if and only if S's safe belief that p is the product of her relevant cognitive abilities (such that her safe cognitive success is to a significant degree creditable to her cognitive agency)* (Pritchard *forthcoming*: 20). Again, in (Pritchard 2010*b*: 76) we can read: 'knowledge is safe belief that arises out of the reliable cognitive traits that make up one's cognitive character, such that one's cognitive success is to a significant degree creditable to one's cognitive character'.

[vii] Accordingly, notice that *E, C,* in the above formulation, refer to one's normal environment. To avoid confusion, if I want to indicate that a subject is situated in an environment other than her normal one, I will write '*E', C''*.

[viii] I here say 'usually' because I do not want to exclude in advance active externalism as proposed in contemporary philosophy of mind. Notice, however, that active externalism should be contrasted with Putnam's (1975) and Burge's (1986) passive externalism. For more details and a thoroughgoing exposition and defense of active externalism, see (Clark 2008). Moreover, Pritchard (2010*a*) and (Palermos 2011) have argued for the introduction of active externalism into contemporary epistemology.

[ix] Thanks to Evan Butts (*forthcoming*) for bringing to my attention Millar's

distinction.

[x] Recall that ability has been defined as a *reliable* cognitive process. The gist of the above considerations is that whereas the *possession* of ability depends both on the existence *and* the reliability of the underlying cognitive process relative to the environment it was acquired for, its *manifestation* depends solely on the (actual or theoretical) reliability of the cognitive process within the environment it is being currently employed (i.e., one's regional environment).

[xi] In contrast, Millar would claim that because of the defective background conditions, the relevant abilities are not *possessed*.

[xii] Another way to put the above point could be to consider, as defined in the beginning of the section, that success rates depend on *conditions*. Remember, however, that local facts are such that if *S* now perceives that *p* then *p* is a local fact. What this means is that local environments are very narrow aspects of the world that cannot plausibly pick out any conditions of one's surroundings but only particular details of it. That is to say, conditions, properly understood, do not refer to specificities but rather to ambient features or situations in the world one is embedded in. Put another way, conditions do not usually concern second by second interactions of a subject with minute details of her surroundings. Instead, it would be more plausible to claim that conditions, when they refer to external facts of the world, should rather be identified with the regional facts: the rest of the facts which the subject could have interacted with had she attended something different from what she actually does.

[xiii] Indicatively, see the Stanford Encyclopedia of Philosophy entry on 'Dispositions' by Micahel Fara.

[xiv] Notice that even though mutations do have causes, their appearance and which kind of phenotypes they will give rise to is supposed to be a matter of luck in the sense that they are not induced for a particular purpose.

[xv] See also (D.E Rumelhart, G.E Hinton, and R.J. Williams 1986).

[xvi] Or a relevantly similar one.

[xvii] Consider also Greco's (1999: 23) very instructive example: 'For example suppose that it seems visually to a person that a cat is sleeping on the couch, and on this basis she believes that there is a sleeping cat on the couch. Suppose also that this belief manifests a disposition that the person has, to trust this sort of experience under these sorts of conditions, when motivated to believe the truth. Now, suppose that much less clearly, it seems visually to the person that a mouse has run across the floor. Not being disposed to trust this kind of fleeting experience, the person refrains from believing until further evidence comes in. The fact that the person, properly motivated, is disposed to trust one kind of experience but not the other, constitutes sensitivity on her part that the former is reliable. There is a clear sense in which she takes the former experience to be adequate to her goal of believing the truth, and takes the latter experience not to be. And this is so even if she has no beliefs about her goals, her reliability, or her experience'.

[xviii] Finally what about skepticism? Do the previous considerations offer a way out of the problem of radical skepticism? Unfortunately, they do not. Skepticism is as alive

and kicking as ever. What the above considerations demonstrate, however, is how bad a joke it would be, were we to be the victims of a skeptical scenario. It is not only the case that all of our beliefs could be an illusion produced by an evil demon, or a supercomputer in the service of a crazy scientist, but our potential jokesters must have also produced such a consistently rich environment that contains evidence in support of the idea that we do not need a loving creator (or luck) being responsible for the veridicality of our abilities, because there are evolutionary and developmental *natural* processes that could plausibly bring about the very same result. Going to such an extent so as to deceive us about the nature of our beliefs, would be a very mean joke indeed.

[xviii]Research into the area of this paper was partly supported by the AHRC-funded 'Extended Knowledge' Project, based at the Eidyn research centre, University of Edinburgh.

References

Butts, E. (*forthcoming*), 'Slim is In: An Argument for a Narrow Conception of Abilities in Epistemology, *Journal of Philosophical Research*.

Burge, T. (1986), 'Individualism and Psychology', *Philosophical Review,* 95: 3-45.

Chisholm, R., M. (1977), *The Theory of Knowledge.* Englewood Cliffs, NJ: Prentice Hall.

Churchland, P. S. (1987), 'Epistemology in the Age of Neuroscience', *The Journal of Philosophy*, Vol. 84, No. 10

Clark, A. (2008), *Supersizing The Mind*. Oxford University Press.

Dennett, D. (2009), 'Darwin's Strange Inversion of Reasoning', in *PNAS,* vol. 106.

Dretske, F. (1970), 'Epistemic Operators', *Journal of Philosophy* 67, 1007-23.

Fara, M. (2006), 'Dispositions', *the Stanford Encyclopedia of Philosophy*.

Goldman, A. (1976), 'Discrimination and Perceptual Knowledge', *Journal of Philosophy* 73, 771-91.

Greco, J. (1999), 'Agent Reliabilism', in *Philosophical Perspectives 13: Epistemology,* James Tomberlin (ed.), Atascadero, CA: Ridgeview Press.

—— (2007), 'The Nature of Ability and the Purpose of Knowledge', *Philosophical Issues* 17, 57- 69.

—— (2009), 'Knowledge and Success From Ability', *Philosophical Studies* 142, 17-26.

—— (2010), *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge University Press.

Kallestrup, J. & Pritchard D. H. (*forthcoming*), 'Virtue Epistemology and Epistemic Twin Earth', *The European Journal of Philosophy.*

Nozick, R. (1981), *Philosophical Explanations*. Oxford University Press, Oxford.

Palermos, S.O. (2011). 'Belief-Forming Processes, Extended', *Review of Philosophy and Psychology*, Vol. 2, No. 4, 741-765.

Pritchard, D., Millar, A., Haddock, A. (2010), *The Nature and Value of Knowledge: Three Investigations.* Oxford University Press.

Pritchard, D. H. (2002), 'Resurrecting the Moorean Response to Scepticism', *International Journal of Philosophical Studies* 10, 283-307.

—— (2008), 'Sensitivity, Safety, And Anti-Luck Epistemology'. In *The Oxford Handbook of Skepticism,* (ed.) J. Greco, (Oxford: Oxford University Press).

—— (2009), *Knowledge*. London: Palgrave Macmillan.

—— (2010*a*), 'Cognitive Ability and the Extended Cognition Thesis'*, Synthese,* 175: 133-151.

—— (2010*b*), 'Knowledge and Understanding', in A. Haddock, A. Millar & D. H. Pritchard, *The Nature and Value of Knowledge: Three Investigations*. Oxford: Oxford University Press.

—— (forthcoming), 'Anti-Luck Virtue Epistemology'*, The Journal of Philosophy*

Putnam, H. (1975), 'The Meaning of "Meaning"', in K. Gunderson (ed.), *Language, Mind and Knowledge.* Minneapolis: University of Minnesota Press.

Rumelhart, D. E., Hinton G. E., and Williams R. J., (1986), 'Learning Internal Representations by Error Propagation', in McClelland and Rumelhart, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* . Cambridge, Mass.: MIT Press.

Sosa, E. (1999), 'How to Defeat Opposition to Moore', *Philosophical Perspectives,* 13, 141-54.

—— (2000), 'Skepticism and Contextualism', *Philosophical Issues* 10, 1-18.

—— (2007), *A Virtue Epistemology: Apt Belief and Reflective Knowledge.* Oxford: Clarendon Press.